

# Identifying and understanding road-constrained areas of interest (AOIs) through spatiotemporal taxi GPS data: A case study in New York City

Yunzhe Liu<sup>\*</sup>, Alex Singleton, Daniel Arribas-bel, Meixu Chen

*Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, Roxby Building, Liverpool L69 7ZT, United Kingdom*

## ARTICLE INFO

### Keywords:

Areas of interest  
ST-DBSCAN  
Public transit  
Taxi GPS  
Urban analytics  
Mobility

## ABSTRACT

Urban areas of interest (AOIs) represent areas within the urban environment featuring high levels of public interaction, with their understanding holding utility for a wide range of urban planning applications.

Within this context, our study proposes a novel space-time analytical framework and implements it to the taxi GPS data for the extent of Manhattan, NYC to identify and describe 31 road-constrained AOIs in terms of their spatiotemporal distribution and contextual characteristics. Our analysis captures many important locations, including but not limited to primary transit hubs, famous cultural venues, open spaces, and some other tourist attractions, prominent landmarks, and commercial centres. Moreover, we respectively analyse these AOIs in terms of their dynamics and contexts by performing further clustering analysis, formulating five temporal clusters delineating the dynamic evolution of the AOIs and four contextual clusters representing their salient contextual characteristics.

## 1. Introduction

Urban areas of interest (AOIs) can be broadly defined as areas within an urban environment that attract people's attention, and which are often related to the generalisation of different types of urban economic activity (Hu et al., 2015; Yuan, Zheng, & Xie, 2012a). AOIs are prevalently characterised by metrics describing high levels of public exposure and frequency of demand and are framed within the literature through the use of various terminologies including functionally-critical locations or urban hotspots (Cai, Jiang, Zhou, & Li, 2018; Qin, Zhou, Wu, & Xu, 2017; Zhou, Fang, Thill, Li, & Li, 2015). It has been argued that a set of locations can be considered as an AOI when they involve various types of infrastructure that are of necessity for people's daily life, such as restaurants, primary workplaces, transport hubs, landmarks, entertainments, schools, and universities (Cai et al., 2018; Chen, Arribas-Bel, & Singleton, 2019).

AOIs are also significant for urban transit planning, location-based services, and the management of daily travel since these areas can be utilised to assign higher priority in the allocation of public resources (Hu et al., 2015; Ma, Meng, Xing, & Li, 2019). Due to the wide range of applications of AOIs, successfully identifying and understanding the characteristics of such urban areas could provide a useful reference basis that benefits multiple stakeholders, including but not limited to tourism

management (van der Zee, Bertocchi, & Vanneste, 2020), the identification of social functions (Zhou, Liu, Qian, Chen, & Tao, 2019), urban environmental study (Chen, Arribas-Bel, & Singleton, 2020), urban vitality analysis (Kim, 2018), traffic planning (Alfeo, Cimino, Egidi, Lepri, & Vaglini, 2018), and public transit management (Ni, Huang, Meng, Zhou, & Su, 2019).

A traditional approach to investigate AOIs is primarily dependent on data derived from questionnaire-based methods such as field surveys or travel diaries. However, these approaches are labour-intensive, time-consuming, and error-prone, thus limiting their usefulness and applicability for large geographic areas (Yuan & Raubal, 2012). Following the rapid development and widespread use of location-based technology, large volumes of spatiotemporal data have been being collected either actively or passively, opening up new opportunities to map out and understand urban dynamics and reveal in-depth relationships between the urban fabric and the human mobility (Arribas-Bel, 2014; Qin et al., 2017). Numerous previous studies have implemented data mining techniques on heterogeneous data sources to identify urban AOIs, for instance, check-in data from social media (Chen et al., 2019; Hu et al., 2015; Kuo, Chan, Fan, & Zipf, 2018; Üsküplü, Terzi, & Kartal, 2020), location data from mobile phones (Yang, Zhao, & Lu, 2016), and point of interest (POI) data from commercial location search engines (Xu, Cui, Zhong, & Wang, 2019).

<sup>\*</sup> Corresponding author.

E-mail addresses: [psyliu7@liverpool.ac.uk](mailto:psyliu7@liverpool.ac.uk), [ucesiue@ucl.ac.uk](mailto:ucesiue@ucl.ac.uk) (Y. Liu).

<https://doi.org/10.1016/j.compenvurbsys.2020.101592>

Received 29 July 2020; Received in revised form 2 December 2020; Accepted 28 December 2020

Available online 12 January 2021

0198-9715/© 2020 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Moreover, as a vital component of urban public transit, taxi trip data from GPS-enabled taxis have also been repurposed to define AOIs in many academic studies. For example, [Garcia, Avendaño, and Vaca \(2018\)](#) utilised the origin-destination (OD) matrix extracted from 69 million records of taxi trips in NYC to identify popular taxi drop-off locations. [Keler, Krisp, and Ding \(2020\)](#) investigated commuter-specific destination hotspots located in NYC by using Boro-taxi drop-off GPS points. [Qin et al. \(2017\)](#) applied a spatiotemporal clustering method on the taxi GPS points extracted from taxi trajectory data to detect urban hotspot areas in Wuhan. [Cai et al. \(2018\)](#) explored urban hotspots and computed their attractiveness index score through utilising one-week of taxi GPS trajectory data collected from 6599 taxis in Kunming.

According to the related studies (see [Cai et al., 2018](#); [Chen et al., 2019](#); [Hu et al., 2015](#); [Kuo et al., 2018](#)), a typical bottom-up AOI detection framework can be summarised as comprising the following three phases:

1. the hotspot detection phase: identifying point clouds (i.e. the AOI prototype) through a density-based clustering method such as DBSCAN;
2. the boundary-defining phase: constructing closed polygons to define the AOI boundary;
3. the analysis phase: clarifying and exploring the characteristics of AOIs.

However, there are several aspects of these phases that require further consideration and improvement. Firstly, the hotspot detection phase is often limited to attributes in 2D planar space, which overwhelmingly concentrate on answering the question of ‘where’ but somewhat ignore the dynamic variation from the temporal aspect of AOIs. Given the fact that not every area in the urban environment is continuously recognised as a hotspot that attracts people’s interest across all time periods, the omission of the temporal dimension may impose challenges in distinguishing different AOIs. For instance, office buildings and transport nodes (e.g. railway stations and airports) are defined as urban AOIs since they are both characterised by overall high traffic volume. However, the overall high traffic volume in the former AOI is more likely to be limited to two peak-time periods of commuting (i.e. morning and evening peak), whereas the latter AOI has a large traffic volume all day except at closing time.

A second research gap relates to those methods used in the boundary-defining phase. It is common to use a set of closed polygons to represent AOI geometrically, since using polygons can “provide simple and accessible representations for areas compared with clustered points” ([Hu et al., 2015](#), 241). Many studies defined the border of an AOI by enclosing identified hotspots through convex hull or bounding box algorithms ([Cai et al., 2018](#); [Hollenstein & Purves, 2010](#)). Although such methods are computationally efficient and convenient to apply, those polygons constructed through convex hulls are very likely to cover superfluous empty areas ([Akdag, Eick, & Chen, 2014](#)). Other studies utilised the concave hull algorithms to define AOI boundaries, such as chi-shape algorithm ([Hu et al., 2015](#)) or alpha-shape ([Chen et al., 2019](#); [Kuo et al., 2018](#)). However, concave hull algorithms are highly susceptible to parameter selection (e.g.,  $\lambda$  in chi-shape and  $\alpha$  in alpha shapes), which is embodied in small changes in parameter settings can make a significant difference in the shape of the calculated polygon ([Chen et al., 2019](#)). Since there is no authoritative guidance on how to obtain the optimal parameters, parameter selection is relatively subjective and can affect the quality of the results returned. Additionally, the feasibility of using polygons to represent AOIs remains to be discussed further, as such geometry only takes the impacts of human activities at AOIs into consideration, while the reshaping influences of urban structure on AOIs are neglected ([Ma et al., 2019](#)).

The third research gap relates to the analysis phase of the three-phase framework. After AOIs are identified, most existing studies mainly

concentrate on their spatial distribution and morphology, but seldom do they explore those latent attributes, in terms of dynamic and contextual aspect, affecting the configuration and characterisation of an AOI. Such circumstance emerges more commonly in studies using traffic data (e.g., taxi GPS) as inputs since there is usually no extra information facilitating further in-depth analysis other than spatiotemporal coordinates.

The unique contribution of this study is the proposal of an enhanced three-phase analytical framework that improves on the aforementioned workflow within the context of a taxi GPS dataset collected for the case study area, i.e. New York City. These methodological enhancements aim to provide new substantive insight into the spatiotemporal dynamics and contextual characteristics of urban AOIs within the New York City and specifically for the Manhattan area. Firstly, we present urban AOIs as both a spatial and temporal phenomenon, implementing the ST-DBSCAN algorithm to detect spatiotemporal taxi trip hotspots. Secondly, in the process of defining the boundary of AOIs, the detected hotspots are linked to road geometry rather than enclosing them with polygons, formulating road-constrained AOIs. Finally, after the construction of AOIs, we utilise the H-K-mean clustering algorithm to conduct an in-depth analysis of these areas in respect of their spatiotemporal dynamics. Additionally, we extract several contextual variables from external open data sources and investigate the salient multidimensional characteristics of the identified AOIs through a geo-demographic analysis.

The remainder of this paper proceeds as follows. [Section 2](#) presents an overview of the case study area and the data used in this study, accompanying with a brief description introducing the main points of the data pre-processing and sampling. [Section 3](#) provides a detailed explanation of the proposed three-phase analytical framework, ranging from essential theoretical context and algorithm introduction to detailed parameter settings and variable selection. [Section 4](#) and its sub-sections respectively depict the results generated from each phase of the proposed framework, which is then followed by a summary of the work and a discussion of future directions in the context of known limitations.

## 2. Data and exploratory analysis

New York City (NYC) is the selected case study area. It is the most densely populated city within the US, with an estimated 8.4 million population distributed over a land area of approximately 784 km<sup>2</sup> ([US Census Bureau, 2019](#)). NYC is situated in the south-east of the state of New York on the US eastern seaboard, including five boroughs: Brooklyn, Queens, Manhattan, Bronx, and Staten Island. Across this area, the New York City Taxi and Limousine Commission (TLC), founded in 1971, is the agency responsible for licensing and regulating all segments of the taxi-related industry, primarily involving Medallion taxis (Yellow taxis), Street Hail Liveries (Green taxis), and For-Hire Vehicles (FHV). In 2018, there were more than 300,000 TLC licensed vehicles servicing across the boroughs of NYC ([TLC, 2018](#)).

Data used in this study were extracted from the TLC database,<sup>1</sup> involving taxi trip records jointly generated by both Yellow taxis and Green taxis in the whole year of 2015. The primary reason we used this 2015 dataset is that it is the latest and most accessible taxi trip data containing detailed GPS coordinates delineating individual taxi travels. Due to privacy issues, since the latter half of the year 2016, the TLC has replaced the provision of original taxi GPS coordinates by aggregating them into designated Taxi Zones,<sup>2</sup> accordingly causing difficulties in analysing them through a density-based algorithm. It should also be mentioned that, although FHV are occupying more and more proportions of taxi trips over the recent years (see [TLC, 2018](#)), trip record data from FHV were excluded from this study since FHV only began submitting trip records in Taxi Zone format after April 2015.

<sup>1</sup> <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

<sup>2</sup> <https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddge>

Data cleaning eliminated taxi trip records that were erroneous or out of bounds, such as GPS coordinates located outside of the study area or too far away from the nearest road network ( $\geq 50$  m); drop-off times that were earlier than pickup times; and unrealistic passenger counts. After the cleaning process, 150,134,156 taxi trip records were retained of the approximately 160 million original trips. Fig. 1 is a hexagon-binning map showing the spatial distribution of the retained taxi trip points. The majority of the NYC taxi trips (approximately 84% of the total taxi GPS points) are found within the Manhattan area, and as such, we subset the data to only focus on this area. Taking computational capacity into consideration, 1% of the samples (i.e. 1,190,646 taxi trips; 2,381,292 pickup and drop-off points), randomly selected from the pre-processed dataset, were subsequently inputted to the follow-up analysis.

The choice of a 1% random sample mirrors previous studies aiming to represent general human mobility patterns (González, Hidalgo, & Barabási, 2008). However, to ensure the validity/stability of findings, multiple 1% random samples of the source taxi GPS data were iteratively selected and tested within our framework to examine the stability of the results. Specifically, we conducted an experiment in which 1% samples of the taxi GPS data were randomly selected multiple times, formulating several testing datasets. Then we examined the output results generated by inputting each of the testing datasets into the first two phases of our framework (introduced in Section 3). On the basis of this iterative experiment, we only retained the AOIs that can be identified every single run, assuring their stability and representativeness, and utilised them to carry out further investigations (i.e. the third phase).

### 3. Methodology framework

Fig. 2 presents a conceptual diagram illustrating an overview of the methodological framework proposed in this study. The framework consists of three phases, generally mirroring the conventional workflow mentioned in Section 1, but containing a methodological enhancement in each phase. Firstly, in the hotspot detection phase, we apply the ST-DBSCAN algorithm to the pre-processed taxi GPS data to detect the spatiotemporal hotspots of the taxi trips located in the case study area. The second phase is boundary-defining, which is responsible for converting the detected taxi hotspots into road-constrained AOIs through the K-Nearest Neighbour (KNN) algorithm that aggregates point clusters to their nearest road segments. The last phase of the framework is the analysis phase, which is comprised by two layers, i.e. dynamic layer and contextual layer, concentrating on extracting knowledge about the dynamic features and the contextual characteristics of the identified AOIs through clustering analysis that is carried out by using hierarchical k-means (H-K-means) algorithm. The remaining subsections respectively describe each phase of our proposed framework in more depth.

#### 3.1. The hotspot detection phase

DBSCAN (density-based spatial clustering for applications with noise) is a commonly applied density-based clustering algorithm for hotspot detection (Ester, Kriegl, Sander, & Xu, 1996), which is configured by two parameters: Epsilon (Eps), the search radius based on a user-defined distance measure, and MinPts, the minimum points within the Eps radius. These parameters jointly determine a minimum density threshold. Point clusters are constructed at locations in which

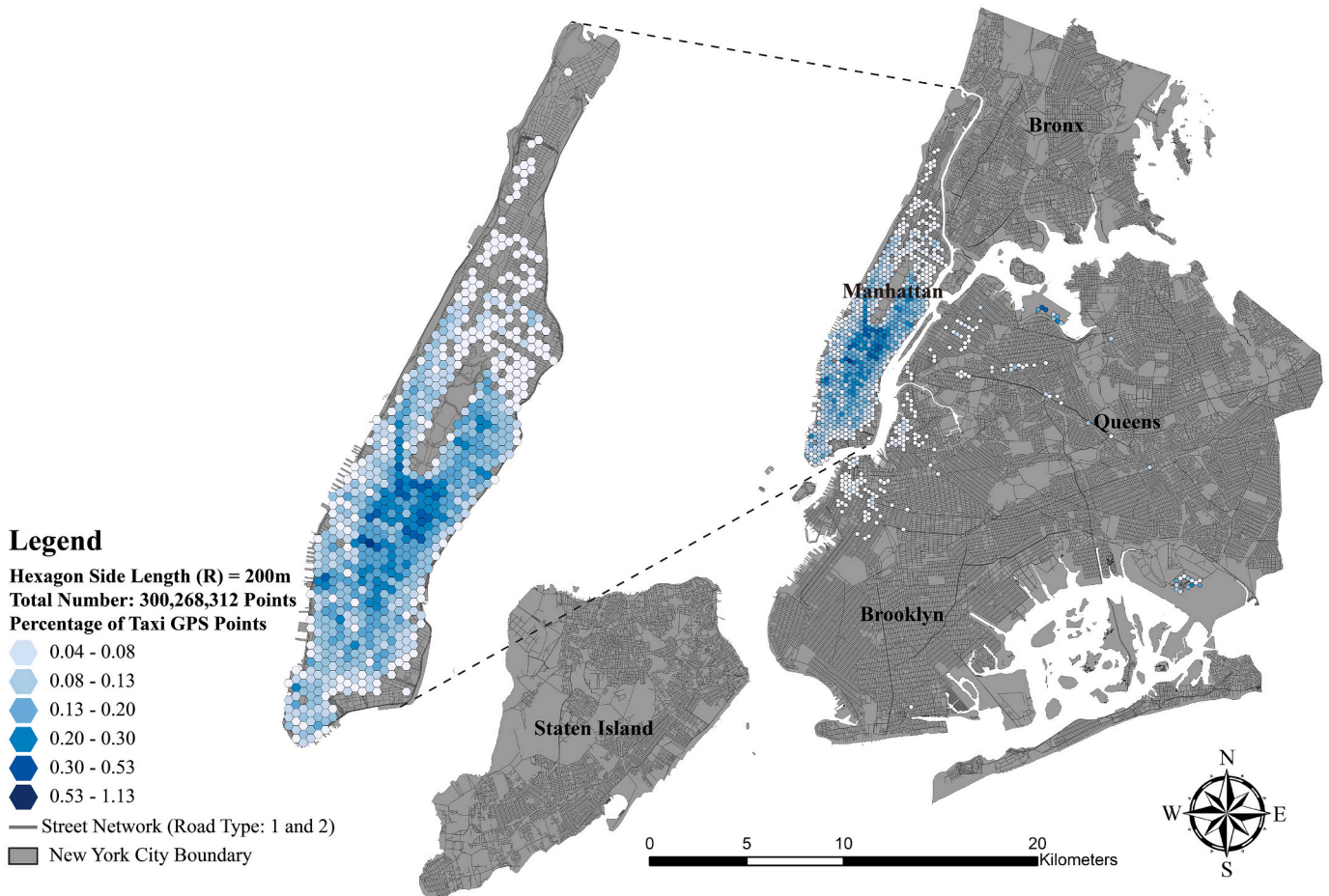


Fig. 1. Spatial distribution of hexagon-binning for pre-processed taxi GPS points in NYC, 2015.



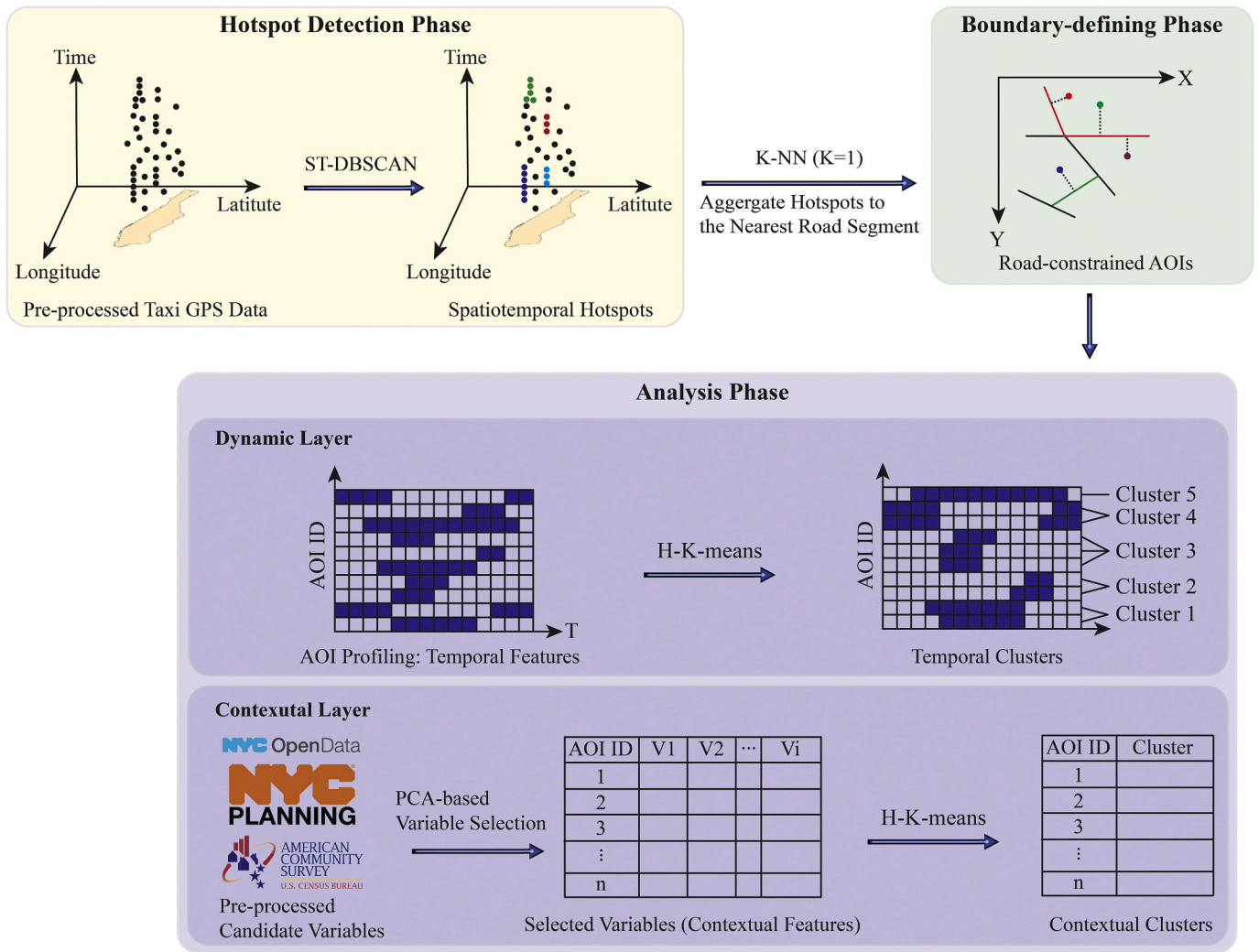


Fig. 2. Conceptual diagram of the proposed analytical framework.

the point density exceeds the specified threshold.

Given the advantages in distinguishing between outliers and clustered points through a relatively simple parameter setting, DBSCAN and its extensional algorithms have been widely employed by many studies to detect hotspots from large-scale geo-referenced data. For instance, Xu et al. (2019) applied DBSCAN to POI data extracted from the Baidu map API to identify the spatial agglomeration of POI-forming functional regions within Wuhan. Zhang, Chen, Wang, and Guan (2016) applied Grid and Kd-tree DBSCAN (GD-DBSCAN) algorithms on taxi pickup locations to identify taxi demand hotspots in Shanghai. Chen et al. (2019) implemented Hierarchical-DBSCAN (HDBSCAN) to geotagged photo data from Flickr to capture the dynamic characteristics of urban AOIs in the inner London area.

Due to the nature of DBSCAN, i.e. using only one distance (Eps) to measure similarity, DBSCAN and most of the abovementioned DBSCAN-based algorithms merely consider spatial attributes in the process of detecting hotspots, resulting in the omission of temporal attributes (Birant & Kut, 2007). However, the urban environment is a complex and constantly changing system, involving various components with multifaceted relationships and interactions (Batty, 2013). Such complexities can be reflected in the changeable type, intensity and distribution of urban resources at different times and locations, referring to both urban dynamics and human mobility (Song, Xia, Jin, Hui, & Li, 2019). From the perspective of urban AOI, not all areas of the urban environment can be recognised as a hotspot over all time periods (Chen et al., 2019; Hu

et al., 2015). We argue here that in many other studies that exclude a temporal dimension, this leads to the capture of only a partial representation of urban AOIs, hence, limiting our understanding of urban functions and their underlying spatiotemporal dynamics.

In order to consider spatial and temporal dimensions simultaneously, ST-DBSCAN (Spatial-temporal Density-Based Spatial Clustering of Applications with Noise), a modified extension of the traditional DBSCAN designed to analyse spatiotemporal data (Birant & Kut, 2007; Shi & Pun-Cheng, 2019), was employed to detect taxi hotspots. Generally, the primary convenience of ST-DBSCAN is that it can identify spatiotemporal clusters with arbitrary shape and noise points (Cheng, Haworth, Anbaroglu, Tanaksaranond, & Wang, 2014). More specifically, according to Birant and Kut (2007), ST-DBSCAN surpasses normal DBSCAN in terms of the three following advantages: firstly, it provides cluster discoverability according to the non-spatial, spatial, and temporal values of objects; secondly, it can effectively detect noise points even when various cluster densities exist; thirdly, it improves clustering quality even if clusters are adjacent to each other. Numerous studies have highlighted the utility of ST-DBSCAN for handling complex spatiotemporal data and the application to many areas of research (see Chen, Bowers, Cheng, Zhang, & Chen, 2020; Iliopoulou, Milioti, Vlahogianni, & Kepaptsoglou, 2020; Shen & Cheng, 2016).

In common with other DBSCAN-based algorithms, ST-DBSCAN also requires predefined parameters before application. According to Birant and Kut (2007), MinPts can be determined by a heuristic method (Eq.



(1)).

$$\text{MinPts} \approx \ln(n) \quad (1)$$

$n$  indicates the total number of observations. In this study, the observations are the 2,381,292 taxi GPS points located in the Manhattan area, NYC. MinPts is accordingly equal to 15.

To define Eps (i.e. Eps1), a  $k$ -distance graph (Fig. 3) delineates ascendingly sorted distances to the  $k$ -nearest neighbours for each object (where  $k = \text{MinPts}$ ). An appropriate Eps value can be selected from the “first valley” of the graph (Birant & Kut, 2007, 214), where there is “an obvious and abrupt change” (Shi & Pun-Cheng, 2019, 7). For this case, we selected 70 m as the Eps value based on this heuristic method.

In addition to MinPt and Eps, Birant and Kut (2007) introduced a second epsilon parameter, i.e. Eps2, to define the search radius for the temporal dimension. Similar to Eps1 mentioned above, a larger value for Eps2 results in broader clusters, while a smaller value generates narrower clusters, delineating a finer temporal resolution. Here we set Eps2 equal to 0.25, representing a 15-min search radius. The primary reason for choosing this temporal resolution was approximately referenced by the average taxi trip time (i.e. 14.8 min), along with the consideration of a convenient result display and interpretation.

### 3.2. The boundary-defining phase

As discussed in Section 1, it is typical in the delineation of urban AOI use an enclosed polygon to define the boundary of the identified point clusters (i.e. hotspots) to formulate AOIs. However, there are growing appeals for alternative representations. Firstly, despite convex-hull and concave-hull algorithms being commonly used in many related studies, both have drawn criticism. The former is sometimes challenged for creating redundant empty areas (Akdag et al., 2014), whilst the latter is susceptible to the choice of parameters, thus involving high subjectivity (Cai et al., 2018; Chen et al., 2019; Hu et al., 2015). Secondly, it can be argued that defining AOI's boundary using enclosed polygons fails to appropriately account for the potential impacts of urban morphology on shaping AOIs since they “only considered the distribution characteristics of data that capture human activities” (Ma et al., 2019, 2). Thirdly, because of the uncertainties caused by inevitable measurement error of GPS, offset between the observed location and the actual location may be a feature of the data inputs: although vehicle GPS location should align with the road network (Yang & Gidófalvi, 2018). Taking such concerns into account, we argue that the road network is a more organic carrier of the detected point clusters, which therefore can be employed to define the boundary of urban AOIs, particularly for an application utilising taxi data since they bound these patterns of mobility (Ma et al., 2019; Yuan, Zheng, & Xie, 2012b).

After projecting the detected taxi trip hotspots onto the 2D plane containing the road network, a KNN algorithm was adopted to aggregate these points to their nearest road segment, formulating road-constrained AOIs. It should be mentioned that, if the road segments are topologically connected, they are considered as one AOI, ensuring that there are no overlapping AOIs.

### 3.3. The analysis phase

After AOIs are identified, most existing studies mainly concentrate on their spatial distribution or temporal evolution pattern, but seldom explore the latent attributes affecting the configuration of AOIs. Such circumstance emerges more commonly in studies using traffic data (e.g. taxi GPS) as inputs since there is no adequate information facilitating further analysis other than spatiotemporal coordinates (i.e. longitude, latitude, and time).

This phase consists of two layers, i.e. dynamic layer and contextual layer, which are designed to extract useful information about the detected AOIs through further clustering analysis from both dynamic

and contextual perspectives. The clustering results generated through each layer will be presented and discussed in Section 4.

#### 3.3.1. The dynamic layer

Since the spatiotemporal hotspots were aggregated to street segments to form the road-constrained AOIs, each AOI can be regarded as proportionally containing at least one or more point clusters over a temporal sequence. Such temporal sequences depict various dynamic patterns exhibited by AOIs. Some AOIs, for instance, only appear at a particular time of day, while others have greater longevity.

In this context, the hierarchical  $k$ -means (H-K-means) clustering algorithm was adopted to classify AOIs into groups based on the similarities in their dynamic pattern. H-K-means provides a hybrid of both hierarchical clustering and  $k$ -means clustering and comprises three steps: first agglomerative hierarchical clustering is implemented to the data to create a  $k$  number of clusters; secondly, the centroids (i.e. the mean value) are calculated for each cluster; finally, these computed centroids are used as the centroid initialisation for the  $k$ -means algorithm (Arai & Ridho Barakbah, 2007; B. Chen, Tai, Harrison, & Pan, 2005).

The optimal number of clusters ( $k$ ) is determined by Gap Statistics, introduced by Tibshirani, Walther, and Hastie (2001) (Eq. (2)), which compares the total within-cluster variation for different values of  $k$  with their expected values under “an appropriate null reference distribution of the data” (p. 412).

$$\text{Gap}_n(k) = E_n^* \{ \log(W_k) \} - \log(W_k) \quad (2)$$

$E_n^*$  denotes the expectation under a sample size  $n$  from the reference distribution.  $W_k$  is the pooled within-cluster sum of squares around the cluster means. The estimation of the optimal clusters  $k$  will be the value that maximises  $\text{Gap}_n(k)$ .

The clustering results could portray the picture of ‘urban pulse’ answering questions, such as where AOIs are and when they emerge and disappear.

#### 3.3.2. The contextual layer

As mentioned previously, due to a lack of further detail on journey purpose, it is insufficient to solely use taxi GPS data to understand the characteristics of identified urban AOIs, for example, to explore what specific features of these AOIs attract taxi passengers and further affect their travel behaviour. In order to gain greater insight into the identified AOIs and improve their interpretability, it is helpful to import supplementary data capturing some contextual attributes that potentially influence individual's travel behaviour, as well as to apply the corresponding analytical method to extract meaningful information about the salient characteristics of urban context from these datasets (Liu & Cheng, 2020). In this study, we utilised a geodemographic classification methodology to extract salient contextual characteristics exhibited by each identified AOIs.

Geodemographic classification is an analytical framework that provides categorical summaries of multidimensional socioeconomic, demographic and built environment characteristics for small geographic areas (Singleton, Spielman, & Folch, 2017). The detailed processes to build a geodemographic classification and the advantages of such classification are well documented (see Alexiou, 2016; Harris, Sleight, & Webber, 2005; Leventhal, 2016; Singleton et al., 2017). Geodemographic classification has an expansive and international lineage, with utility for both private and public sectors applications and for various geographic extents (Gale, Singleton, Bates, & Longley, 2016; Singleton & Longley, 2015; Singleton & Spielman, 2014). The implementation of geodemographic classification for this study can be regarded as a bespoke application designed to differentiate urban AOIs in Manhattan, NYC.

Numerous studies have investigated the linkage between urban context and travel behaviour over the past decades (Cervero &

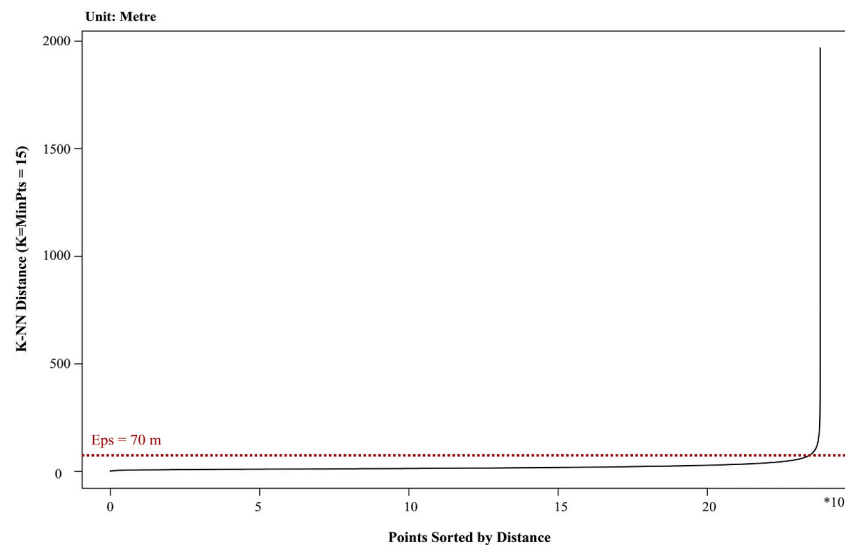


Fig. 3. KNN distance graph ( $K = 15$ ) used to determine Eps (Eps = 70 m).

Kockelman, 1997; Dieleman, Dijst, & Burghouwt, 2002; Ewing & Cervero, 2010; Ma, Mitchell, & Heppenstall, 2014; Pan, Shen, & Zhang, 2009). For instance, Ewing and Cervero (2010) found that an individual's travel mode choice can be influenced by the demographic and socioeconomic characteristics of the household as well as the built environment characteristics of the surrounding area, which provided additional 'D' variables to the well-established 'three Ds' principle (i.e. density, diversity, and design) introduced by Cervero and Kockelman (1997). More recently, Liu, Singleton, and Arribas-Bel (2020) presented a study containing a systematic literature review over 29 contemporary studies related to the impacts of the urban context on people's travel behaviour. They pointed out that although most of the studies still aligned with the 'D' variables, some of the variables they used have beyond the scope of the traditional 'D' variables, implying broader or context-specific considerations. They further categorised those variables into four domains, namely, Land Use and Built Environment (LB), Location and Accessibility (LA), Socioeconomic and Demographic (SD), and Transit-related (T), guiding the variable selection for their research about creating a contextual transit-oriented development (TOD) typology for NYC.

Given the overlapping research context and case study area, we acknowledged the systematic literature review conducted by Liu et al. (2020) and utilised their presented four variable-domains as a reference to guide our initial variable selection. With extra consideration of the availability and consistency of the data (note that the 2015 taxi data were used in this study), 52 candidate variables were initially selected (Table 1), which were extracted from the following four open data sources, i.e. American Community Survey (ACS),<sup>3</sup> NYC Open Data,<sup>4</sup> Smart Location Database (SLD),<sup>5</sup> and NYC Planning.<sup>6</sup>

Inevitably, such a large number of candidate variables and the resulting high dimensionality we argue would lead to harmful effects in the following cluster analysis. Numerous studies have discussed the negative impact caused by the high dimensionality on the clustering performance, which is also known as 'the dimensional curse', including dramatically increasing the demand for computational power and storage capacity, lowering the efficiency of the clustering algorithm, impairing the output interpretability (Iguyon & Elisseff, 2003; Renjith,

Sreekumar, & Jathavedan, 2020; Weber, Schek, & Blott, 1998). Apart from the potential threats from high dimensionality, multicollinearity between the candidate variables is also problematic (Sambandam, 2003). The existence of variable pairs with high correlation is harmful to the clustering performance since such dimensions are effectively assigned more weight during the clustering process (Harris et al., 2005; Sambandam, 2003).

In order to alleviate the adverse impacts of high dimensionality and multicollinearity, we employed a principal component analysis (PCA)-based variable selection framework, proposed by Liu, Singleton, and Arribas-Bel (2019), to "select the smallest possible subset of variables that can represent the main variance within a universe of potential inputs being considered" (Liu et al., 2019, 253). PCA is a feature transformation methods, which has a long history of being applied across multiple disciplines to accomplish dimensionality reduction (Ma et al., 2019; Malhi & Gao, 2004; Webber, 1975). Through linear transformation, PCA finds a set of orthogonal space to maximise the variance in each coordinate axis (Abdi & Williams, 2010), to project high-dimensional data onto a low-dimensional representation, while preserving the original data features as much as possible (Ma et al., 2019). The variable-selection framework proposed by Liu et al. (2019) consists of multiple stages, that not only select variables according to the average contribution of the input variables to the principal components (PCs) but also filters variables based on their correlation between each other. The minimum spanning tree (MST) was integrated into the framework to filter out variable pairs with relatively high correlation (correlation coefficient  $\geq \pm 0.75$ ). Additionally, their framework also considers such impacts on overall clustering quality, which provides additional utility for this study. A full description of the PCA-based variable selection framework, its properties, parameter settings, and relative strengths and weaknesses is beyond the scope of this section however presented by Liu et al. (2019).

Many of the variables related to specific points of interest, and as such were aggregated into the road-constrained AOIs using the KNN algorithm ( $K = 1$ ) that was applied in the boundary-defining phase. Values of some variables, such as Floor area ratio (FAR), were averaged during the aggregation process, whereas others (e.g. many of the ACS variables) were aggregated based up their intersection with the AOI. The last column of Table 1 shows the checklist indicating the contextual variables that were selected after the application of the PCA-based selection method. 27 out of 52 candidate variables were included as inputs.

After the selected variables were assembled for each AOI, the Box-

<sup>3</sup> <https://www.census.gov/programs-surveys/acs/>

<sup>4</sup> <https://opendata.cityofnewyork.us/>

<sup>5</sup> <https://www.epa.gov/smartgrowth/smart-location-mapping>

<sup>6</sup> <https://www1.nyc.gov/site/planning/data-maps/open-data.page>

**Table 1**

Initial 52 candidate variables and selected variables from the PCA-based variable selection framework proposed by Liu et al. (2019).

Data sources	Code	Domain	Variables title	Description	Checklist
ACS	B01001	SD	Age: 0–4	% of population aged between 0 and 4	
		SD	Age: 5–14	% of population aged between 5 and 14	*
		SD	Age: 15–19	% of population aged between 15 and 19	
		SD	Age: 20–24	% of population aged between 20 and 24	
		SD	Age: 25–44	% of population aged between 25 and 44	*
		SD	Age: 45–64	% of population aged between 45 and 64	
		SD	Age: 65 & above	% of population aged 65 and above	*
	B08303	LA	TTtW: <5	% of workers whose travel time to work is less than 5 min	
		LA	TTtW: 5–14	% of workers whose travel time to work is between 5 and 14 min	*
		LA	TTtW: 15–29	% of workers whose travel time to work is between 15 and 29 min	
		LA	TTtW: 30–44	% of workers whose travel time to work is between 30 and 44 min	*
		LA	TTtW: 45–59	% of workers whose travel time to work is between 45 and 59 min	*
		LA	TTtW: >60	% of workers whose travel time to work is longer than 60 min	
	B15003	SD	EA: No school	% of population have no qualifications	*
		SD	EA: Elementary school	% of population attained kindergarten to 5th grade	
		SD	EA: Middle school	% of population attained 6th to 8th grade	
		SD	EA: High school	% of population attained 9th to 12th grade	*
		SD	EA: College/ Bachelor	% of population attained college or bachelor's degree	
		SD	EA: Master/ Doctorate	% of population attained master or doctorate degree	*
	B19013	SD			

**Table 1 (continued)**

Data sources	Code	Domain	Variables title	Description	Checklist
SLD	B24010	SD	Median Income	Household median income in the past 12 months	
			OT: M.B.S. A.	% of workers in management, business, science, and art occupations	*
			OT: S.	% of workers in service occupations	*
			OT: S.O.	% of workers in sales and office occupations	
			OT: N.C.M.	% of workers in natural resources, construction, and maintenance occupations	
	B01003	LB	Population Density	% of workers in production, transportation, and material moving occupations	
			D4a	Distance from the population-weighted centroid to the nearest transit stop (meters)	
			D1c	Gross employment density (jobs/acre)	*
			D4d	Aggregate frequency of transit service per square mile	*
			Intersection Density	Number of street intersections by road length	*
NYCP	STC	LB	Tree Density	Number of street trees by road length	
			Bicycle Facilities	Number of Citi-bike, bicycle routes and parking shelters by road length	*
			Bus Facilities	Number of bus stops by road length	*
			LU: R	% of building & poi categorised as residential use	*
	MapPLUTO	LB	LU: C	% of building & poi categorised as commercial use	*
			LU: TU	% of building & poi categorised as transport and utility	*
			LU: PSCI	% of building & poi categorised as public	*

(continued on next page)



Table 1 (continued)

Data sources	Code	Domain	Variables title	Description	Checklist
		LB	LU: OSR	service and institution % of building & poi categorised as open space	
		LB	LU: V	and recreation % of building & poi categorised as vacant	
		LB	LU: Mixed	% of building & poi categorised as mixed-use	*
		LB	FAR	Floor area ratio (gross floor area/area of plot)	*
		LB	Landmark Density	Number of landmarks by road length	*
		LB	BT: Detached	% of building unit categorised as detached	
		LB	BT: Attached	% of building unit categorised as attached	*
		LB	BT: Semi-Attached	% of building unit categorised as semi-attached	
		LB	BT: Apartment	% of building unit categorised as apartment	
		LB	YB: 2010 / Later	% of building built in 2010 or later	*
		LB	YB: 2000–2009	% of building built between 2000 and 2009	*
		LB	YB: 1980–1999	% of building built between 1989 and 1999	
		LB	YB: 1960–1979	% of building built between 1960 and 1979	
		LB	YB: 1940–1959	% of building built between 1940 and 1959	
		LB	YB: 1939/ Earlier	% of building built in 1939 or later	*

Cox transformation (Box & Cox, 1964) (Eq. (3)) was employed to convert abnormally distributed variables to approximate normality. Furthermore, since the variables are measured on different scales, z-scores (Eq. (4)) were applied as a method of standardisation.

$$x'_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log x_i, & \text{if } \lambda = 0. \end{cases} \quad (3)$$

$x'_i$  is the transformed value;  $\lambda$  ranges from  $-5$  to  $5$ , which can be estimated using the profile likelihood function to achieve 'optimal value'.

$$z_i = \frac{x_i - \mu}{\sigma} \quad (4)$$

$z_i$  is the standardised value,  $x_i$  is an original value,  $\mu$  is the mean of  $x_i$ , and  $\sigma$  is the standard deviation from the mean.

The variables were subsequently clustered through H-K-means, and the Gap Statistics mentioned in Section 3.3.1 were utilised once again to define the optimal number of clusters. The clustering result provides summary measures of the urban context, revealing the salient

characteristics distinguishing AOIs from other urban areas. Furthermore, in order to improve the interpretability of revealed clusters, it is typical to assign shorthand names and written "pen portraits" descriptions for each of the clusters within the built geodemographic classification (Alexiou, 2016; Harris et al., 2005).

## 4. Results

### 4.1. Identified AOIs in Manhattan, NYC

Fig. 4 presents the spatial distribution of the 31 identified urban AOIs. These areas are featured by major transportation hubs, such as the West 39th Street Ferry Terminal (AOI 18), Pennsylvania Station (AOI 15), and Grand Central Station (AOI 16); famous cultural venues, such as the Lincoln Centre for the Performing Arts (AOI 26), the Whitney Museum of American Art (AOI 8), and the Metropolitan Museum of Art (AOI 30); open spaces, such as Central Park (AOI 24) and Union Square (AOI 6); and some other tourist attractions, prominent landmarks, and commercial centres, such as Columbus Circle (AOI 25), the Empire State Building (AOI 13), the Rockefeller Centre (AOI 20), and the One World Trade Centre (AOI 1).

### 4.2. Dynamic features of AOIs

As discussed earlier, an advantage of the ST-DBSCAN algorithm is that in addition to the spatial attributes of the urban AOI, the temporal characteristics are also preserved, enabling further exploration of their dynamic evolution throughout the day. As such, the 31 identified urban AOIs were further classified into five temporal clusters representing different types of dynamic patterns. Fig. 5 contains a sorted heatmap presenting the temporal distribution of the clustering results, followed by a map showing their spatial distribution (Fig. 6). Based on such patterns, furthermore, shorthand names and descriptive profiles were generated for each AOI cluster.

#### 4.2.1. Constant AOIs

Most AOIs classified in this group are located in Midtown of Manhattan, covering various major transit hubs (e.g. Pennsylvania Station, AOI 15), and integrated commercial, retail centres (e.g. Rockefeller Centre, AOI 20). AOIs from this cluster are continuously exposed to a high volume of taxi activity lasting approximately the whole day, and as such is one of the most stable AOIs in Manhattan.

#### 4.2.2. Noon AOIs

AOIs of this group distribute evenly across Manhattan from north to south, with no specific agglomerations. These AOIs record gradually increased taxi flow at around 9:30, a peak at high noon, and a reduction after 17:30, which could be affected by business opening hours.

#### 4.2.3. Morning AOIs

Experiencing high taxi travel demand between 6:00 and 10:30 in the morning, AOIs in this group are primarily identified in areas proximal to major commercial centres (e.g. One World Trade Centre, AOI 1) or public institutions, such as hospitals and medical institutions (e.g. Weill Cornell Medical Centre, AOI 27), which could indicate a typical morning peak commuting pattern.

#### 4.2.4. Late night AOIs

AOIs from this group are mainly identified in south Manhattan. AOIs begin to emerge after 17:30 and continuously attract taxi travels until 3:00 in the early morning of the next day, which might either suggests a recreational pattern reflecting the nightlife in Manhattan or residential-oriented pattern, or combination of both.

#### 4.2.5. Evening AOIs

AOIs of this group are diffuse over Manhattan from Midtown (Union

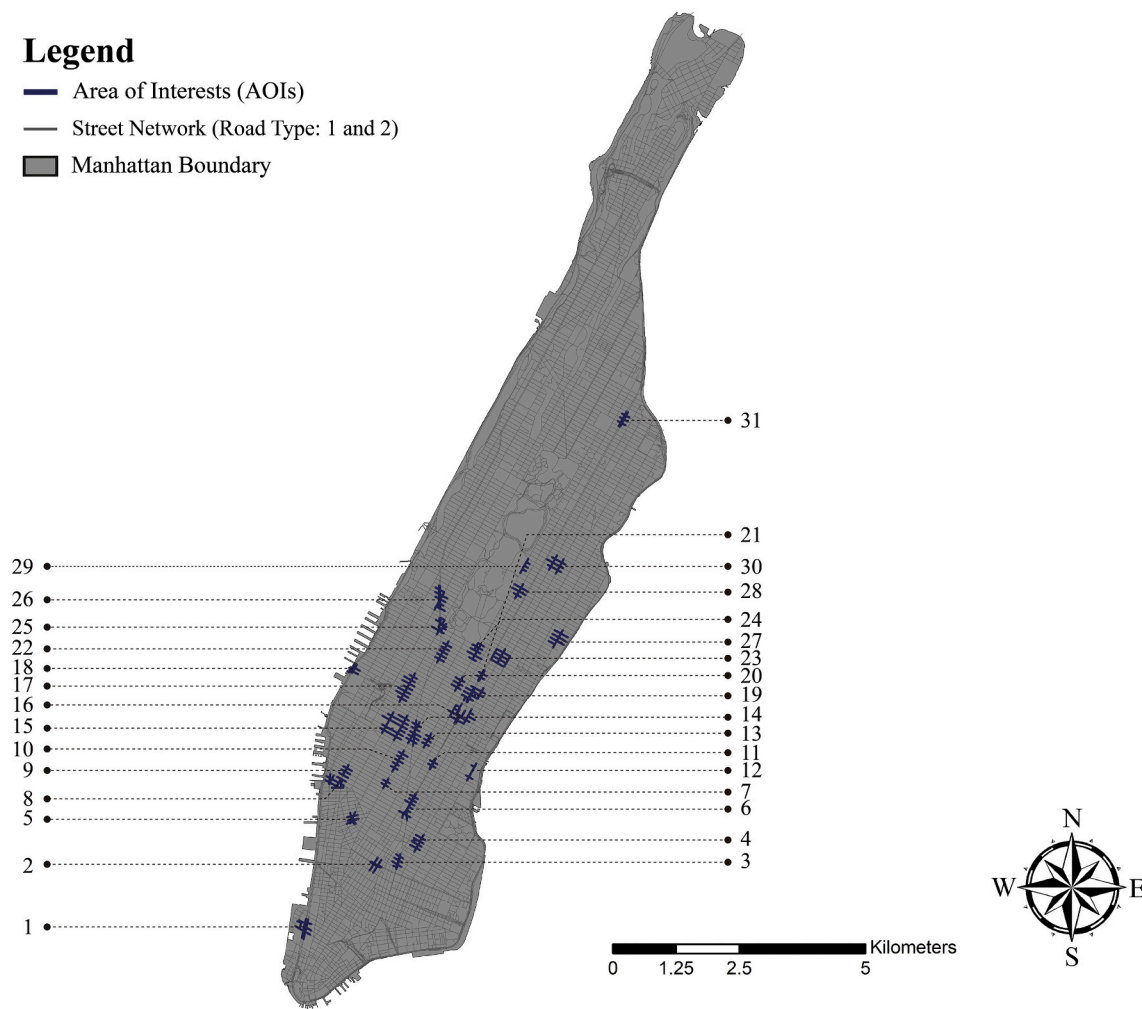


Fig. 4. Geographic distribution of 31 identified AOIs in NYC.

Square, AOI 6) to the Upper West Side (Lincoln Square, AOI 26). These AOIs emerge at around 17:00, peak at around 21:30, and entirely disappear before midnight, indicating an off-peak recreational-oriented travel pattern.

#### 4.3. The contextual feature of AOIs

Fig. 7 presents a map illustrating the spatial distribution of the geodemographic classification that was generated from applying H-K-means to the 27 variables retained by the PCA variable selection. The identified 31 AOIs were classified into four clusters, i.e. Major transit hubs, High-rise integrated commercial, Residential heritage mix, and Public institution mix, delineating four different salient multidimensional characteristics extracted from the contextual variables.

Index scores (i.e.  $x/x^- * 100$ ) were computed for the retained variables and were displayed within each cluster in Fig. 8. These scores reflect the (over-) underrepresentation of a target attribute compared to the average value (i.e. a score of 100). An index score of 50 would be equivalent to a rate that is half the average, and 200 would be double. Using both the map and scores, descriptive profiles were generated.

##### 4.3.1. Major transit hubs

AOIs of this cluster cover primary public transit nodes in Manhattan, predominantly manifested by the high level of transit frequency and the surrounding transport-oriented buildings and facilities. These nodes facilitate inter-/intra city flows, including a ferry terminal (AOI 18),

railway station (AOI 31), and an interstate bus terminal (AOI 17).

##### 4.3.2. High-rise integrated commercial

Commercial-use skyscrapers are very likely to be located in proximity to AOIs from this group since the average floor area ratio is dramatically higher than the average, exemplified by the high-rise office buildings near the One World Trade Centre (AOI 1). These areas are likely to be the leading employment destinations in Manhattan due to the short travel-to-work time and the high level of the job density.

##### 4.3.3. Residential heritage mix

AOIs of this cluster mainly agglomerate in Midtown Manhattan. Areas approximating to these AOIs are likely to contain many old buildings built earlier than 1939 and have had been primarily utilised for residential purposes, while the mixed-use buildings and facilities are also much in evidence (e.g. multipurpose areas near the Pennsylvania Station, AOI 15). Landmark destinations within these AOIs are significantly higher than the regional average, which may be attractive for tourists and travellers.

##### 4.3.4. Public institution mix

These AOIs are prevalently located in Upper Manhattan, although they can be found across Manhattan. Buildings or facilities located near this type of AOIs are likely to be used for many purposes, including residential usages, retailing markets, culture venues, public services (e.g. hospitals), and research or educational institutions.

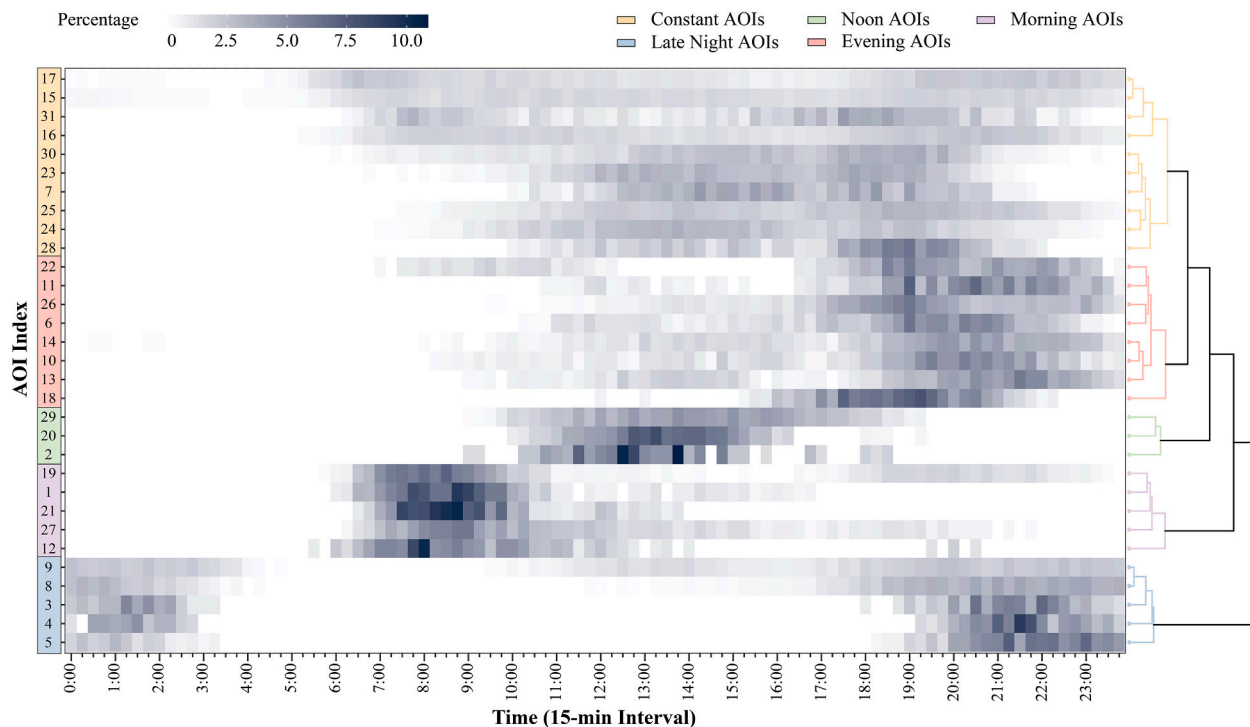


Fig. 5. The temporal distribution of AOIs (by 15-min interval).

#### 4.4. Integrated spatiotemporal dynamics and context

The main objective of this study was to understand how AOIs are represented both from contextual and spatiotemporal perspectives. Accordingly, the intersection of the temporal and contextual classifications was analysed through cross-tabulation, and the result presented in Fig. 9. The heatmap illustrates the frequency and proportion of AOIs categorised at the intersection of the two typologies. The result indicates a general correspondence between the two classifications with some emerging differences.

As the major gateways of NYC and interchange platforms facilitating multimodal inter-/intra-city journeys, two out of three AOIs from the 'Major transit hubs' unsurprisingly correspond to the 'Constant AOIs' featuring consistent exposure to high volumes of taxi traffic throughout the day. It should be noticed that although the areas near the ferry terminal (i.e. AOI 18) are also classified as 'Major transit hubs', these areas are only recognised as an AOI after 16.30 (i.e. Evening AOIs), which might indicate a typical evening return peak use.

The intersection also reveals regular commuting patterns. Nearly 60% of those AOIs classified as 'High-rise integrated commercial' are respectively occupied by 'Morning AOIs' and 'Evening AOIs', manifesting a typical bimodal commuting pattern. However, there is also correspondence between the AOIs categorised as 'Residential heritage mix' and 'Late Night AOIs', suggesting a residential-oriented function.

Moreover, characterised by mixed and compact land use, AOIs from the 'High-rise integrated commercial' and 'Public institution mix' categories present various temporal usage patterns, which with more diffuse representation over the four temporal clusters, with the exception of 'Late Night AOIs'. Such a pattern reflects a wide variety of essential roles in people's daily life, which could satisfy multiple demands, including entertainment, public services, commuting, shopping, tourism and other aspects.

## 5. Discussion and conclusions

The measurement and ascription of urban AOIs are of continued

interest within the field of urban mobility studies. The wide availability of large-scale spatiotemporal data has enabled a variety of new methods of identifying and understanding urban AOIs through the application of density-based cluster analysis, which can generally be conceptualised into a framework comprising three phases: hotspot detection, boundary-defining, and analysis. We identified how such frameworks as those currently implemented contain several limitations across each phase. Firstly, due to the nature of the traditional DBSCAN algorithm, many of the existing studies overwhelmingly concentrated on the spatial aspect of the AOI, while a more integrated view combining spatial and temporal dimensions was somewhat overlooked. Secondly, using enclosed polygon to define the boundary of AOI from those identified hotspot clusters may not form the most appropriate units for analysis given that they lack the attributes of the underlying urban morphology that may inform the identified patterns. Finally, after AOIs are identified, most existing studies neglect the characterisation of those latent attributes affecting the formation of AOIs.

Within this context, our study proposed a new analytical framework that is guided by a conventional three-phase workflow, yet addressed the abovementioned research. The ST-DBSCAN algorithm was employed as the core of the first phase to detect spatiotemporal hotspots. In the second phase, the road network was used to define the boundary of urban AOI; and finally, the dynamic features and contextual features of urban AOI were exposed and investigated. The proposed framework was applied to a taxi GPS dataset extracted from the selected case study area, New York City.

Our enhanced framework identified 31 unique AOIs across the spatial extent of Manhattan. Most of the AOI locations were highly correlated to famous places, such as landmarks, culture venues, open spaces, commercial centres, and transit stations. The spatiotemporal dynamics of the extracted AOIs were considered through further cluster analysis conducted using the H-K-means algorithm. The 31 detected AOIs were classified into five unique clusters (i.e. Temporal Clusters), respectively, representing different types of spatiotemporal activity. Furthermore, the contextual features of AOIs were considered by importing 52 candidate variables from supplementary open data portals.



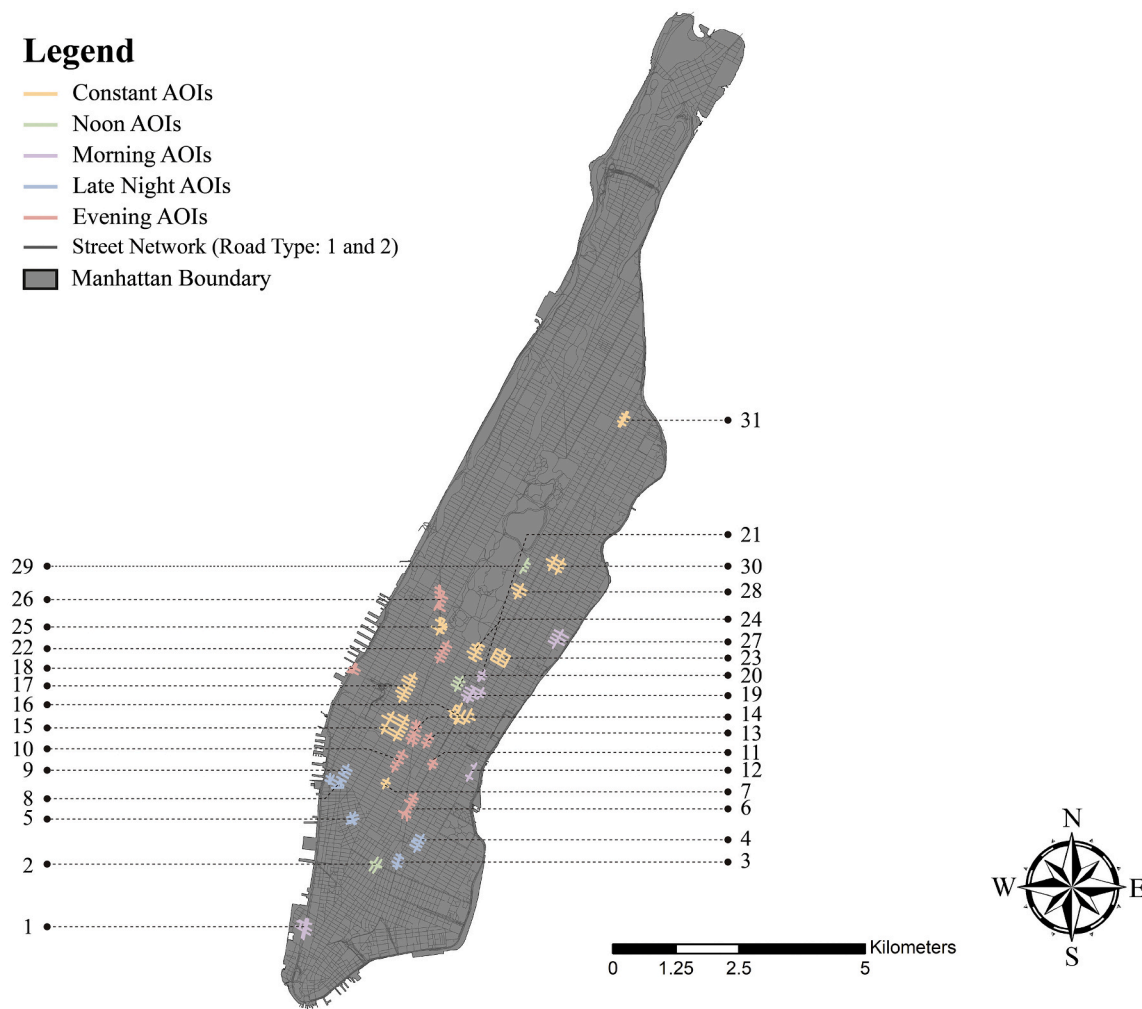


Fig. 6. Geographic distribution of five temporal clusters.

A PCA-based variable selection framework proposed by Liu et al. (2019) was employed to filter out redundant variables, which eventually retained 27 variables that identified five salient AOI clusters (i.e. Contextual Clusters). These clusters were named, described, and mapped. Through cross-tabulating the abovementioned two types of AOI clusters, a high degree of correspondence was found, reflecting the interrelation between the context and dynamics of AOIs.

The utility of defining road-constrained AOIs alongside their dynamic and contextual characteristics we envisage will benefit multiple stakeholders. For urban planners and policymakers, they are more likely to identify urban areas with greater priority and issue more context-based policies, assisting in allocating limited urban resources more effectively. For transport agencies and operators, enhanced spatiotemporal information about the urban AOIs could help to mitigate traffic congestions and provide timely adjustment to the provision of public transport. For taxi drivers, enhanced knowledge of trip hotspots will assist in making more purposeful route selections to maximise the potential for passenger demand. For tourists and travellers, the identified urban AOIs might be utilised as an informative city guide; and for retailers and business managers, our results could assist them with site selection and targeted advertising.

One limitation of this study relates to the parameter selection of ST-DBSCAN. The method used in this study to define MinPt and Eps is primarily based on the heuristic method suggested by the Birant and Kut (2007), which requires further justification in terms of practical application. In another context, Chen et al. (2019) suggested using 1% of the

observations to define the MinPt in their study on the detection of urban AOIs in London. In our case, however, if 1% of the observations were employed to define the parameter, the algorithm would fail to identify any clusters since the MinPt is too large (i.e. MinPt is more than 2000). As we discussed previously, there are no standard rules guiding the parameter selection, meaning that the parameter setting may be adjusted according to the actual conditions. As such, we envisage further work looking at optimised methods for parameter selection. Nonetheless, despite such caveat, this paper has presented an innovative methodological framework to identify and understand urban AOIs in terms of both context and dynamics, and will likely be a useful framework for applications within other urban contexts.

The presented approach is extendable in many ways. One direction of future work that would be favourable to the quality of value of the outcomes is the integration with the other emerging datasets. Since the landscape of the traditional taxi market has been changing by the rapid rise of 'ride-hailing' businesses such as Uber and Lyft, a growing number of taxi travellers replace their traditional on-street-hailing with more convenient app-hailing (NYDOT, 2018; Willis & Tranos, 2021). In this context, it is possible to either compare the spatiotemporal differences between the urban AOIs formed by the traditional taxi GPS data and those formed by the app-based for-hire vehicle data; or integrate them together to deepen our understandings about the urban AOIs more comprehensively within the context of the current taxi market. Furthermore, with more public transit datasets are becoming publicly available, it is possible to identify and compare AOIs through using data

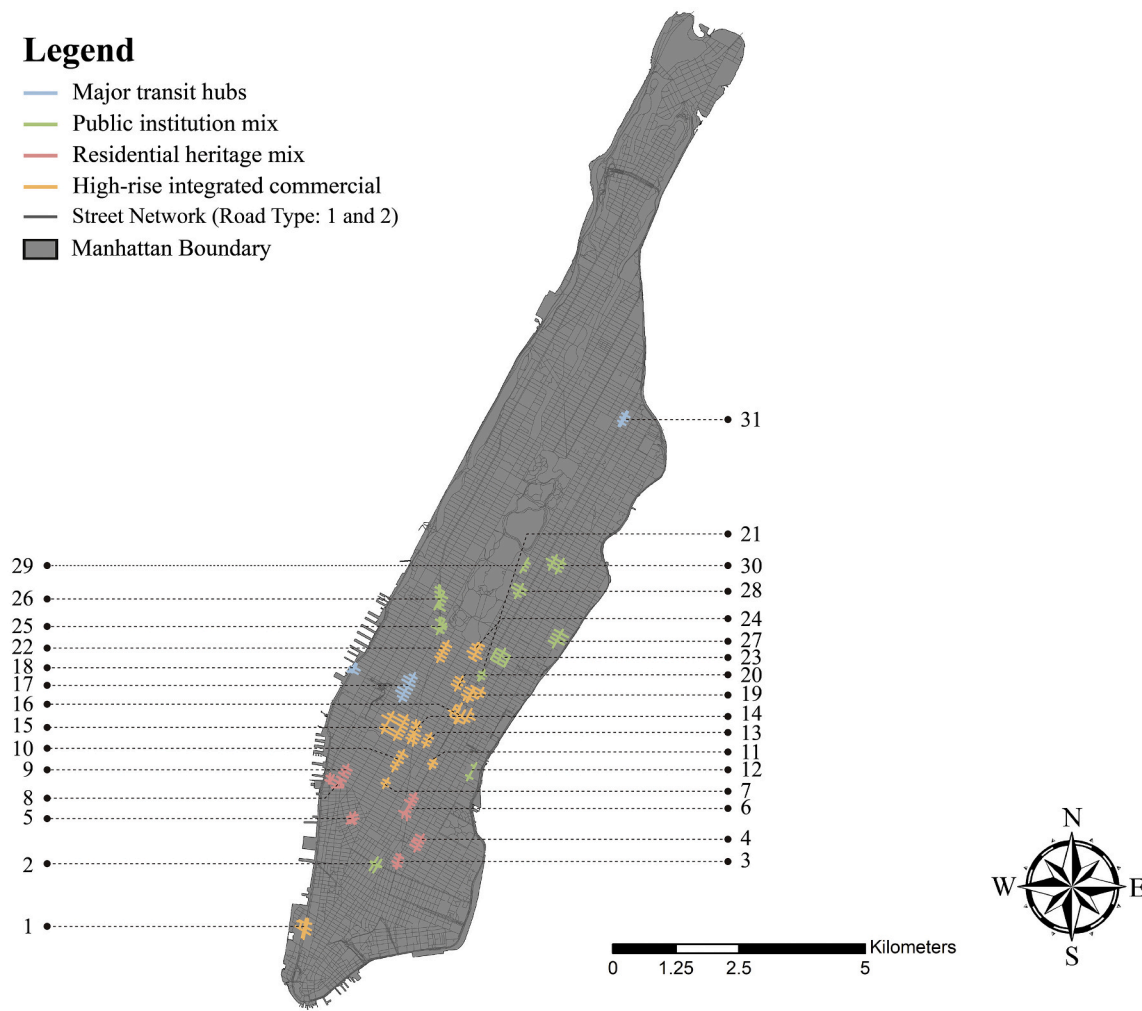


Fig. 7. Geographic distribution of four contextual clusters.

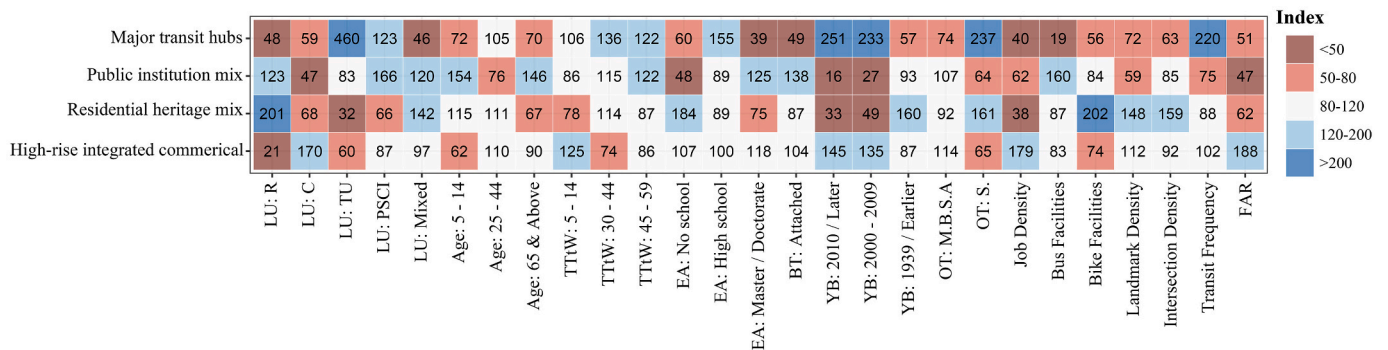


Fig. 8. Index scores by four contextual clusters.

from other travel modes, which might demonstrate manifold differences of interest between multimodal travellers, e.g. active mobility and motorised road users (Keler et al., 2020).

#### Authorship statement

Authorship contributions:

- Conception and design of study: Yunzhe Liu;
- Acquisition of data: Yunzhe Liu, Meixu Chen;

- Analysis and/or interpretation of data: Yunzhe Liu
- Drafting the manuscript: Yunzhe Liu;
- Revising the manuscript critically for important intellectual content: Yunzhe Liu, Alex Singleton, Daniel Arribas-bel, Meixu Chen
- Approval of the version of the manuscript to be published: Yunzhe Liu, Alex Singleton, Daniel Arribas-bel, Meixu Chen

#### Funding

This research did not receive any specific grant from funding

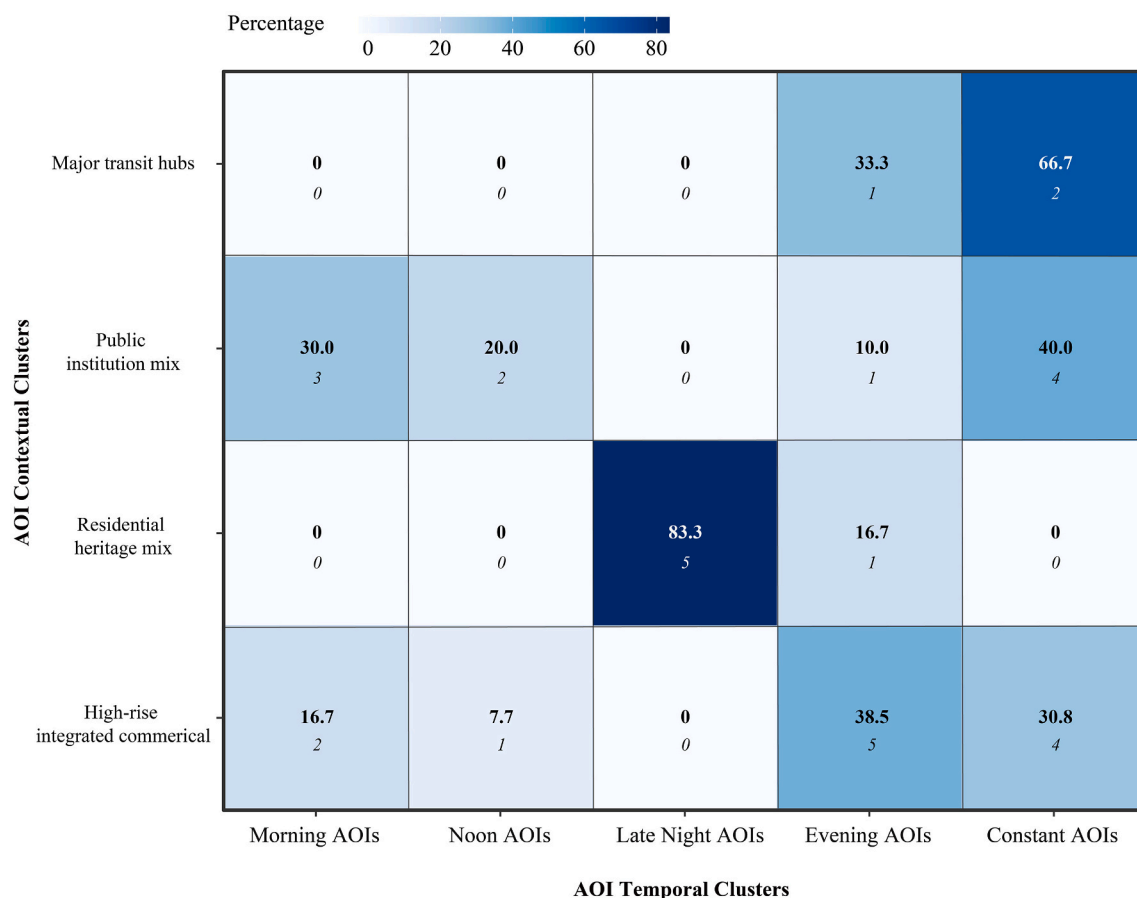


Fig. 9. Cross-tabulation: AOI frequency and percentage by Contextual Clusters and Temporal Clusters in Manhattan. Italic number shows the actual number of AOIs; Bold number shows the percentage.

agencies in the public, commercial, or not-for-profit sectors.

#### Disclosure statement

No potential conflict of interest was reported by the authors.

#### References

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. In *Wiley interdisciplinary reviews: Computational statistics*. <https://doi.org/10.1002/wics.101>.
- Akdag, F., Eick, C. F., & Chen, G. (2014). *Creating polygon models for spatial clusters* (pp. 493–499). Cham: Springer. [https://doi.org/10.1007/978-3-319-08326-1\\_50](https://doi.org/10.1007/978-3-319-08326-1_50).
- Alexiou, A. (2016). Putting “Geo” into Geodemographics: Evaluating the performance of national classification systems within regional contexts. <https://livrepository.liverpool.ac.uk/3007463/>.
- Alfeo, A. L., Cimino, M. G. C. A., Egidi, S., Lepri, B., & Vaglini, G. (2018). A stigmergy-based analysis of city hotspots to discover trends and anomalies in urban transportation usage. *IEEE Transactions on Intelligent Transportation Systems*. <https://doi.org/10.1109/TITS.2018.2817558>.
- Arai, K., & Ridho Barakbah, A. (2007). *Hierarchical K-means: An algorithm for centroids initialisation for K-means* (In Rep. Fac. Sci. Engrg. Reports of the Faculty of Science and Engineering).
- Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*. <https://doi.org/10.1016/j.apgeog.2013.09.012>.
- Batty, M. (2013). *The new science of cities*. MIT Press. <https://mitpress.mit.edu/books/new-science-cities>.
- Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*. <https://doi.org/10.1016/j.datak.2006.01.013>.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B: Methodological*. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>.
- Cai, L., Jiang, F., Zhou, W., & Li, K. (2018). Design and application of an attractiveness index for urban hotspots based on GPS trajectory data. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2018.2869434>.
- Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*. [https://doi.org/10.1016/S1361-9209\(97\)00009-6](https://doi.org/10.1016/S1361-9209(97)00009-6).
- Chen, B., Tai, P. C., Harrison, R., & Pan, Y. (2005). Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis. In *2005 IEEE computational systems bioinformatics conference, workshops and poster abstracts*. <https://doi.org/10.1109/CSBW.2005.98>.
- Chen, M., Arribas-Bel, D., & Singleton, A. (2019). Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems*. <https://doi.org/10.1007/s10109-018-0284-3>.
- Chen, M., Arribas-Bel, D., & Singleton, A. (2020). Quantifying the characteristics of the local urban environment through geotagged flickr photographs and image recognition. *ISPRS International Journal of Geo-Information*. <https://doi.org/10.3390/ijgi9040264>.
- Chen, T., Bowers, K., Cheng, T., Zhang, Y., & Chen, P. (2020). Exploring the homogeneity of theft offenders in spatio-temporal crime hotspots. *Crime Science*, 9(1), 9. <https://doi.org/10.1186/s40163-020-00115-8>.
- Cheng, T., Haworth, J., Anbaroglu, B., Tanaksaranond, G., & Wang, J. (2014). Spatiotemporal data mining. In *Handbook of regional science*. [https://doi.org/10.1007/978-3-642-23430-9\\_68](https://doi.org/10.1007/978-3-642-23430-9_68).
- Dieleman, F. M., Dijst, M., & Burghouwt, G. (2002). Urban form and travel behaviour: Micro-level household attributes and residential context. *Urban Studies*. <https://doi.org/10.1080/00420980220112801>.
- Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD '96* (pp. 226–231). <https://doi.org/10.5555/3001460.3001507>.
- Ewing, R., & Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association*. <https://doi.org/10.1080/01944361003766766>.
- Gale, C., Singleton, A., Bates, A., & Longley, P. (2016). Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science*, 12 (2016), 1–27. <https://doi.org/10.5311/JOSIS.2016.12.232>.
- Garcia, J. C., Avendaño, A., & Vaca, C. (2018). *Where to go in Brooklyn: NYC mobility patterns from taxi rides* (pp. 203–212). Cham: Springer. [https://doi.org/10.1007/978-3-319-77703-0\\_20](https://doi.org/10.1007/978-3-319-77703-0_20).
- González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature*. <https://doi.org/10.1038/nature06958>.
- Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting*. John Wiley & Sons Ltd. <https://www.wiley.com/en-gb/Geodemographics+GIS+and+Neighbourhood+Targeting-p-9780470864135>.



- Hollenstein, L., & Purves, R. S. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*. <https://doi.org/10.5311/JOSIS.2010.1.3>.
- Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*. <https://doi.org/10.1016/j.compenvurbsys.2015.09.001>.
- Iguyon, L., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*. <https://doi.org/10.1162/153244303322753616>.
- Iliopoulou, C. A., Milioti, C. P., Vlahogianni, E. I., & Kepaptsoglou, K. L. (2020). Identifying spatio-temporal patterns of bus bunching in urban networks. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. <https://doi.org/10.1080/15472450.2020.1722949>.
- Keler, A., Krisp, J. M., & Ding, L. (2020). Extracting commuter-specific destination hotspots from trip destination data—comparing the boro taxi service with Citi Bike in NYC. *Geo-Spatial Information Science*. <https://doi.org/10.1080/10095020.2019.1621008>.
- Kim, Y. L. (2018). Seoul's Wi-Fi hotspots: Wi-Fi access points as an indicator of urban vitality. *Computers, Environment and Urban Systems*. <https://doi.org/10.1016/j.compenvurbsys.2018.06.004>.
- Kuo, C. L., Chan, T. C., Fan, I. C., & Zipf, A. (2018). Efficient method for POI/ROI discovery using Flickr geotagged photos. *ISPRS International Journal of Geo-Information*. <https://doi.org/10.3390/ijgi7030121>.
- Leventhal, B. (2016). *Geodemographics for marketers: Using location analysis for research and marketing*. Kogan Page.
- Liu, Y., & Cheng, T. (2020). Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science*. <https://doi.org/10.1080/23249935.2018.1493549>.
- Liu, Y., Singleton, A., & Arribas-Bel, D. (2019). A Principal Component Analysis (PCA)-based framework for automated variable selection in geodemographic classification. *Geo-Spatial Information Science*, 22(4), 251–264. <https://doi.org/10.1080/10095020.2019.1621549>.
- Liu, Y., Singleton, A., & Arribas-Bel, D. (2020). Considering context and dynamics: A classification of transit-orientated development for New York City. *Journal of Transport Geography*. <https://doi.org/10.1016/j.jtrangeo.2020.102711>.
- Ma, H., Meng, Y., Xing, H., & Li, C. (2019). Investigating road-constrained spatial distributions and semantic attractiveness for area of interest. *Sustainability (Switzerland)*. <https://doi.org/10.3390/su11174624>.
- Ma, J., Mitchell, G., & Heppenstall, A. (2014). Daily travel behaviour in Beijing, China. An analysis of workers' trip chains, and the role of socio-demographics and urban form. *Habitat International*. <https://doi.org/10.1016/j.habitatint.2014.04.008>.
- Malhi, A., & Gao, R. X. (2004). PCA-based feature selection scheme for machine defect classification. *IEEE Transactions on Instrumentation and Measurement*. <https://doi.org/10.1109/TIM.2004.834070>.
- Ni, X., Huang, H., Meng, Y., Zhou, S., & Su, B. (2019). An urban road-traffic commuting dynamics study based on hotspot clustering and a new proposed urban commuting electrostatics model. *ISPRS International Journal of Geo-Information*. <https://doi.org/10.3390/ijgi8040190>.
- NYDOT. (2018). NYC mobility report (White Paper, June, 7–8) <http://www.nyc.gov/html/dot/downloads/pdf/mobility-report-2018-print.pdf>.
- Pan, H., Shen, Q., & Zhang, M. (2009). Influence of urban form on travel behaviour in four neighbourhoods of Shanghai. *Urban Studies*. <https://doi.org/10.1177/0042098008099355>.
- Qin, K., Zhou, Q., Wu, T., & Xu, Y. Q. (2017). Hotspots detection from trajectory data based on spatiotemporal data field clustering. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. <https://doi.org/10.5194/isprs-archives-XLII-2-W7-1319-2017>.
- Renjith, S., Sreekumar, A., & Jathavedan, M. (2020). *Pragmatic evaluation of the impact of dimensionality reduction in the performance of clustering algorithms* (pp. 499–512). Singapore: Springer. [https://doi.org/10.1007/978-981-15-5558-9\\_45](https://doi.org/10.1007/978-981-15-5558-9_45).
- Sambandam, R. (2003). Cluster analysis gets complicated. In *Marketing research*.
- Shen, J., & Cheng, T. (2016). A framework for identifying activity groups from individual space-time profiles. *International Journal of Geographical Information Science*. <https://doi.org/10.1080/13658816.2016.1139119>.
- Shi, Z., & Pun-Cheng, L. S. C. (2019). Spatiotemporal data clustering: A survey of methods. *ISPRS International Journal of Geo-Information*. <https://doi.org/10.3390/ijgi8030112>.
- Singleton, A., & Longley, P. (2015). The internal structure of Greater London: A comparison of national and regional geodemographic models. *Geo: Geography and Environment*. <https://doi.org/10.1002/geo2.7>.
- Singleton, A., & Spielman, S. (2014). The past, present, and future of geodemographic research in the United States and United Kingdom. *The Professional Geographer*, 66(4), 558–567. <https://doi.org/10.1080/00330124.2013.848764>.
- Singleton, A., Spielman, S., & Folch, D. (2017). *Urban analytics*. SAGE Publication Ltd. <https://uk.sagepub.com/en-gb/eur/urban-analytics/book249267>.
- Song, S., Xia, T., Jin, D., Hui, P., & Li, Y. (2019). *UrbanRhythm: Revealing urban dynamics hidden in mobility data* (In arXiv).
- Taxi and Limousine Commission. (2018). 2018 fact book. [https://www1.nyc.gov/assets/tlc/downloads/pdf/2018\\_tlc\\_factbook.pdf](https://www1.nyc.gov/assets/tlc/downloads/pdf/2018_tlc_factbook.pdf).
- Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. <https://doi.org/10.1111/1467-9868.00293>.
- United States Census Bureau. (2019). Quick facts. <https://www.census.gov/quickfacts/newyorkcitynewyork>.
- Üsküplü, T., Terzi, F., & Kartal, H. (2020). Discovering activity patterns in the City by social media network data: A case study of Istanbul. *Applied Spatial Analysis and Policy*. <https://doi.org/10.1007/s12061-020-09336-5>.
- Webber, R. (1975). Liverpool social area study 1971 data: Final report. In *PRAG technical paper 14*. Planning Research Applications Group, Centre for Environmental Studies. <https://catalogue.nla.gov.au/Record/746774>.
- Weber, R., Schek, H. J., & Blott, S. (1998). A similarity-search analysis methods and performance study for in high-dimensional spaces. In *Proceedings of the 24th VLDB conference* (pp. 1–8).
- Willis, G., & Tranos, E. (2021). Using 'Big Data' to understand the impacts of Uber on taxis in New York City. *Travel Behaviour and Society*, 22, 94–107. <https://doi.org/10.1016/j.tbs.2020.08.003>.
- Xu, Z., Cui, G., Zhong, M., & Wang, X. (2019). Anomalous urban mobility pattern detection based on GPS trajectories and POI data. *ISPRS International Journal of Geo-Information*. <https://doi.org/10.3390/ijgi8070308>.
- Yang, C., & Gidófalvi, G. (2018). Mining and visual exploration of closed contiguous sequential patterns in trajectories. *International Journal of Geographical Information Science*, 32(7), 1282–1304. <https://doi.org/10.1080/13658816.2017.1393542>.
- Yang, X., Zhao, Z., & Lu, S. (2016). Exploring spatial-temporal patterns of urban human mobility hotspots. *Sustainability (Switzerland)*. <https://doi.org/10.3390/su8070674>.
- Yuan, J., Zheng, Y., & Xie, X. (2012a). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*. <https://doi.org/10.1145/2339530.2339561>.
- Yuan, N. J., Zheng, Y., & Xie, X. (2012b). Segmentation of urban areas using road networks. Msr-Tr-2012-65 <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/mapsegmentation.pdf>.
- Yuan, Y., & Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. In N. Xiao, M. Kwan, M. Goodchild, & S. Shekhar (Eds.), *Geographic information science. GIScience 2012. Lecture notes in computer science* (pp. 354–367). Springer.
- van der Zee, E., Bertocchi, D., & Vanneste, D. (2020). Distribution of tourists within urban heritage destinations: A hot spot/cold spot analysis of TripAdvisor data as support for destination management. *Current Issues in Tourism*. <https://doi.org/10.1080/13683500.2018.1491955>.
- Zhang, L., Chen, C., Wang, Y., & Guan, X. (2016). Exploiting taxi demand hotspots based on vehicular big data analytics. In *2016 IEEE 84th vehicular technology conference (VTC-Fall)* (pp. 1–5). <https://doi.org/10.1109/VTCFall.2016.7881010>.
- Zhou, T., Liu, X., Qian, Z., Chen, H., & Tao, F. (2019). Automatic identification of the social functions of areas of interest (AOIs) using the standard hour-day-spectrum approach. *ISPRS International Journal of Geo-Information*, 9(1), 7. <https://doi.org/10.3390/ijgi9010007>.
- Zhou, Y., Fang, Z., Thill, J. C., Li, Q., & Li, Y. (2015). Functionally critical locations in an urban transportation network: Identification and space-time analysis using taxi trajectories. *Computers, Environment and Urban Systems*. <https://doi.org/10.1016/j.compenvurbsys.2015.03.001>.

**Yunzhe Liu** is a Ph.D. student in Geographic Data Science Lab at University of Liverpool. Before that he was graduated (with Distinction) from the MSc Geographic Information Sciences at University College London. His research interests focus on urban analytics, human mobility, geodemographics, spatiotemporal data mining with application in transport, and urban planning.

**Alex Singleton** is a professor of Geographic Information Science at the University of Liverpool, Deputy Director of the ESRC Consumer Data Research Centre (CDRC) and Director of the ESRC Data Analytics & Society CDT. His research is concerned with how the complexities of individual behaviours, attitudes and contexts manifest spatially, and can be represented and understood through a framework of geographic data science.

**Daniel Arribas-bel** is a senior lecturer in Geographic Data Science at the Department of Geography and Planning, and member of the Geographic Data Science Lab, at the University of Liverpool (UK). He is also an ESRC Fellow at the Alan Turing Institute. His research focuses on Geographic data science and new forms of data, urban economics and regional science, and open source scientific computing and reproducibility.

**Meixu Chen** is a Ph.D. student in Geographic Data Science Lab at University of Liverpool. Her research focuses on urban analytics, urban areas of interest, GeoAI, image recognition, social media, and housing market analysis.